Coding for Data Science and Data Management
Module of Data Management

MongoDB

Stefano Montanelli
Department of Computer Science
Università degli Studi di Milano
stefano.montanelli@unimi.it

Essentials

- MongoDB (name derived from «hu**mongo**us») is a cross-platform, document-oriented database engine
- MongoDB has been firstly released in 2007 and it progressively gained popularity in the framework of database solutions

http://db-engines.com

Essentials

- MongoDB is made up of databases which contain collections
  - A collection shares enough in common with the notion of table in a relational database
- A collection is made up of documents
  - A document shares enough in common with the notion of tuple/record in a relational database

Essentials

- Each document is made up of fields
  - A field shares enough in common with the notion of attribute in a relational database
- Collections can be indexed, which improves lookup and sorting performance
  - Indexes in MongoDB function mostly like their RDBMS counterparts.

Serialization format

- MongoDB employs BSON (Binary JSON – Binary Javascript Object Notation)
- BSON is a binary serialization format used
  - to store documents
  - to enforce remote procedure calls
- The MongoDB syntax is **case sensitive**

Primary keys

- Each document has a predifined **_id** field which represents the primary key of the document within the collection
- The value of the _id field is unique inside the collection
- If not inserted, _id is automatically generated to provide a unique ObjectId

References

- References across different documents (also when belonging to different databases) can be represented in two ways
- **Normalized way**. The field value of a document contains the value of the _id field of the referenced object (similar to the notion of foreign key in relational databases)

References

- References across different documents (also when belonging to different databases) can be represented in two ways
- **Denormalized way**. The field value of a document contains the entire referenced object (data embedding)

Useful commands of Mongo DB

- show databases
  - List the available dbs stored on the instance
- use *namedb*
  - Connect to the *namdb* database
- db.getCollectionNames()
  - List the available collections of the current database
  - The reserved keyword «db» is used to reference the current database

MongoDB hands on

- For everything else, refer to
  «the Little MongoDB Book»

- Alternative source of information:
  The offical website of MongoDB manual

The aggregation pipeline

- «*The aggregation pipeline is a framework for data aggregation modeled on the concept of data processing pipelines. Documents enter a multi-stage pipeline that transforms the documents into aggregated results*»
- https://docs.mongodb.com/manual/core/aggregation-pipeline/

Pipeline

- The MongoDB aggregation pipeline consists of stages
- Each stage transforms the documents as they pass through the pipeline
- Pipeline stages do not need to produce one output document for every input document
  - e.g., some stages may generate new documents or filter out documents

Aggregation stages

- The pipeline is equipped with a method **db.*collectioname*.aggregate()**
- The stages in the pipeline are passed to the aggregate method in sequence as they appear
- The output of a stage is the input of the subsequent one

- https://docs.mongodb.com/manual/reference/operator/aggregation-pipeline/#aggregation-pipeline-operator-reference